REMARKS/ARGUMENTS

The amendment is in response to the Office Action dated April 22, 2004. Claims 1-23 are pending in the present application. Applicant has amended claims 1-6, 9-14 and 17, 18, 20 and 21. Accordingly, claims 1-23 remain pending in the present application.

Amended Claims

Independent claims 1, 9, and 17 were amended to incorporate a portion of claim 2, 10 and 18, respectively. Claims 2, 10 and 18 were amended to delete the portions incorporated into the base claims. Claims 3-6, 11-14, 20 and 21 were amended to correct claim dependencies. No new matter has been presented.

Claim Rejections

The Examiner rejected claims 1-23 under 35 U.S.C. §103(a) as being unpatentable over Meyerzon et al. (U.S. Patent No. 6,631,369) in view of Nelson et al. (U.S. Patent No. 6,243,713). In so doing, the Examiner stated:

> Regarding claims 1 and 9, Meyerzon discloses a method for retrieving information using a search engine comprising the steps of:
>
> (a) retrieving a document to be indexed (see col. 4, lines 43-54, Meyerzon);
>
> (b) generating a document extract corresponding to the document (see col. 4, lines 53-67); and
>
> (d) storing the plurality of tokens in a search index, wherein the search engine accesses the search index to retrieve information in one or more document extracts satisfying a search query (see col. 7, lines 44-65 and col. 8, lines 1-10, Meyerzon).
>
> Meyerzon, however, does not explicitly disclose (c) decomposing the document extract into a plurality of tokens. Nelson, on the other hand, discloses the retrieval system for retrieval of multimedia information including the decomposing the document into a plurality of tokens (see abstract of Nelson). It would have been obvious to one of ordinary skill in the art at the time of the

invention to modify Meyerzon to include the claimed feature as taught by Nelson .

. . .

Regarding claim 17, Meyerzon discloses a system for retrieving information, wherein the system includes a search engine comprising:

- means for retrieving a document from a documentary repository (see col. 4, lines 43-54 and element 200, Fig. 2 and corresponding text, Meyerzon);
- an information extractor coupled to the means for retrieving, wherein the information extractor generates a document extract corresponding to the document (see col. 4, lines 53-67, Meyerzon). Each document is retrieved form the web site process and the data is extracted from each of these retrieved documents. Therefore, there must be an extractor for the extracting process;
- a storage device (100, Fig. 2 and corresponding text, Meyerzon) coupled to the information extractor for storing the document extract;
- a search engine indexer (300, Fig. 2) coupled to the storage device; and
- a search index (400, Fig. 2) coupled to the search engine indexer to retrieve information in one or more document extracts satisfying a search query (see col. 7, lines 44-65 and col. 8, lines 1-10; Fig. 2 and corresponding text, Meyerzon).

Meyerzon, however, does not explicitly disclose the step of decomposing the document extract into a plurality of tokens. Nelson, on the other hand, discloses the retrieval system for retrieval of multimedia information including the decomposing the document into a plurality of tokens (see abstract of Nelson). It would have been obvious to one of ordinary skill in the art at the time of the invention to modify Meyerzon to include the claimed feature as taught by Nelson .

. . .

Regarding claims 2, 10 and 18, Meyerzon/Nelson combination further discloses the steps of (b1) extracting a portion of the document that characterizes the document's subject content to form the document extract; and (b2) recording positional information of the portion extracted within the document (see col. 6, lines 1-10, Nelson).

Applicant respectfully traverses.

The present invention relates to retrieving relevant data in large collections of documents. According to the present invention, a search index that reflects the characteristic portions of a document is created by utilizing an information extractor, which examines a document and generates a document extract. The document extract comprises only a portion of the document that is most characteristic of the document as a whole. Thus, the search index is based on the document extract, and not on the document itself.

Through aspects of the present invention, the search index is far more refined in its content because it does not contain references to inconsequential portions of a document. Moreover the size of the search index is greatly reduced because only a portion of the document is parsed. This, in turn, allows the search process to proceed more rapidly because less information is analyzed.

The present invention, as recited in claim 1 provides:

> 1. A method for retrieving information using a search engine comprising the steps of:
>
> (a) retrieving a document to be indexed;
>
> (b) generating a document extract corresponding to the document by extracting a portion of the document that characterizes the document's subject content to form the document extract;
>
> (c) decomposing the document extract into a plurality of tokens; and
>
> (d) storing the plurality of tokens in a search index, wherein the search engine accesses the search index to retrieve information in one or more document extracts satisfying a search query.

Independent claims 9 and 17 are computer readable medium and system claims, respectively, having scopes similar to that of claim 1.

Independent claims 1, 9 and 17 are Allowable.

Applicant respectfully submits that none of the cited references, alone or in combination, teach or suggest "extracting a portion of the document that characterizes the document's subject content to form the document extract," as recited in claims 1, 9 and 17. As stated above, the content of the search index is based on the document extracts which characterize the subject content of the corresponding documents. Accordingly, the search index is based on the *semantic value* of the documents, as opposed to just the words or components of the document.

In contrast, Meyerzon, is directed to minimizing the number of requests a web crawler makes to a document server to obtain the "increment" of the document set relative to the set of

documents received during the previous crawl. Nelson is directed to indexing compound documents in a unified common index. In Nelson, a compound document, i.e., a document containing multimedia components, is broken up into its constituent components (e.g., text, audio, images) and one or more tokens is created for each component. The components and their tokens are then stored in the unified common index (col. 2, lines 19-27).

While Meyerzon teaches "extracting the data from each of these retrieved documents and storing the data in an index" (column 4, lines 55-59), and Nelson teaches decomposing the compound document into its constituent multimedia components, indexing the components, and storing the indexed data in an index (column 5, lines 52-67), neither reference focuses on building an index based on the document's subject content. In particular, neither Meyerzon nor Nelson, singularly or in combination, teach or suggest "extracting *a portion of the document that characterizes the document's subject content* to form the document extract" corresponding to the document, as recited in claims 1, 9 and 17.

In the Office Action, the Examiner asserts that Meyerzon/Nelson teaches this feature at column 6, lines 1-10 of Nelson. Applicant disagrees. The cited portion of Nelson states that the index comprises a set of tokens that represent some aspect of a multimedia component in the compound document. The token "also has additional reference data that defines at least the position in the compound document of the original multimedia component (or portion thereof) that is associated with the token, and may include the actual, or preferably, processed data extracted from, and representative of the original component." (Column 6, lines 1-10). Nothing in the cited portion teaches or suggests "extracting a portion of the document *that characterizes the document's subject content* to form the document extract" corresponding to the document, as recited in claims 1, 9 and 17.

For the reasons presented above, Applicant respectfully submits that the cited references fail to teach or suggest the cooperation of elements recited in claims 1, 9 and 17 and that those claims are therefore allowable over the cited references. Claims 2-8, 10-16 and 18-23 depend on claims 1, 9 and 17, respectively, and the arguments above apply with full force to claims 2-8, 10-16 and 18-23. Accordingly, Applicant respectfully submits that claims 2-8, 10-16 and 18-23 are also allowable over the cited references.

Dependent Claims 5-7, 13-15 and 21 are Allowable for Alternative Reasons

Applicant respectfully submits that dependent claims 5-7, 13-15 and 21 are allowable over the cited references for reasons in addition to being dependent on allowable base claims. First, neither reference teaches or suggests "extracting from the document a collection of sentences that are characteristic of the document's subject content to form a document summary," as recited in claims 5 and 13. In the Office Action, the Examiner states that Nelson teaches this feature at column 6, lines 1-34. That portion, however, discusses tokens and how that are generated. It mentions that "a text component (e.g., a paragraph of text) may be indexed by a number of tokens, each representing one or more words of the text component" (col. 6, lines 10-13), and that "a text token in most cases will represent an actual text string; e.g., the token 'house' will be used to index the word 'house.'" (Col. 6, lines 17-19). Nothing in Nelson teaches or suggests "extracting from the document a collection of sentences *that are characteristic of the document's subject content* to form a document summary," as recited in claims 5 and 13.

Second, neither reference teaches or suggests "selecting from the document extract one of a whole sentence, a portion of a sentence, a word, and a feature," as recited in claims 6 and 14. As discussed above, neither reference teaches or suggests generating the document extract.

Therefore, it follows that neither reference can teach or suggest selecting any portion or part of the document extract. In the Office Action, the Examiner states that Nelson teaches this feature at column 6, lines 1-34. Nevertheless, as discussed above, Applicant respectfully submits that the cited portion makes no mention or suggestion of "selecting from the document extract one of a whole sentence, a portion of a sentence, a word, and a feature," as recited in claims 6 and 14.

Finally, neither reference teaches or suggests "selecting [the plurality of tokens] based on frequency of occurrence, word-salient-measure, proximity to the beginning of a paragraph, proximity the beginning of the document, and proximity to and position within a heading and a caption," as recited in claims 7, 15 and 21. The Examiner states that Nelson teaches this feature at column 19, lines 39-51. That portion of Nelson, however, discusses creating query structures for execution during retrieval. In this process, the tokens are combined together with query operators to form the query structure. Two groups of query operators are used: mathematical operators and proximity operators. Applicant respectfully submits that nothing in Nelson teaches or suggests "selecting [the plurality of tokens] based on frequency of occurrence, word-salient-measure, proximity to the beginning of a paragraph, proximity the beginning of the document, and proximity to and position within a heading and a caption," as recited in claims 7, 15 and 21.

Conclusion

In view of the foregoing, Applicant submits that claims 1-23 are allowable over the cited references. Applicant respectfully requests reconsideration and allowance of the claims as now presented.

Applicant's attorney believes that this application is in condition for allowance. Should any unresolved issues remain, Examiner is invited to call Applicant's attorney at the telephone number indicated below.

Respectfully submitted,

SAWYER LAW GROUP LLP

July 20, 2004
Date

Joyce Tom
Attorney for Applicant
Reg. No. 48,681
(650) 493-4540